INTRASOFT INTERNATIONAL

EU IoT

TINYML: STATE OF THE ART AND RESEARCH CHALLENGES

John Soldatos, Intrasoft International

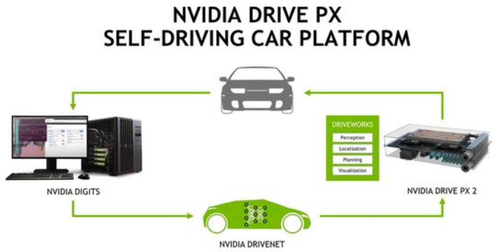
NGIOT IoT & Edge II Workshop, On-Line, December 7th, 2020

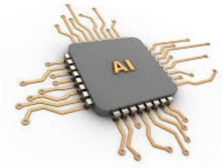
TINYML

Background and Motivation

- Is AI/ML in the Cloud Always the Best Option?
  - Siri, Alexa, OK Google etc.: benefit from an instant response
  - Industrial maintenance: More timely detection of failures and abnormalies
  - Agriculture: Instant disease detection using a plant’s image
  - In several cases ML/AI «at the edge» dramatically reduces costs & complexity, and limits potential data privacy leaks.
- TinyML: An Alternative Form of Machine Learning and AI at the Far Edge

Cloud AI (32 GB, TFLOPS/s) → Smart Phone AI (4GB, GFLOPs/s) → TinyML (<500KB, MFLOPs/s)





LEVERAGE THE BENEFITS OF THE FAR EDGE

TinyML & Edge Computing Benefits

Low Latency

- No need to send data to the cloud – excellent for operations close to the field

Reduced Power Consumption

- Microcontrollers are power-efficient and can operate for a long time without being charged

Improved Environmental Performance

- A step to Green AI – Cloud AI can have very poor environmental performance

Bandwidth Savings

- Data are not transferred to the server – low bandwidth operation

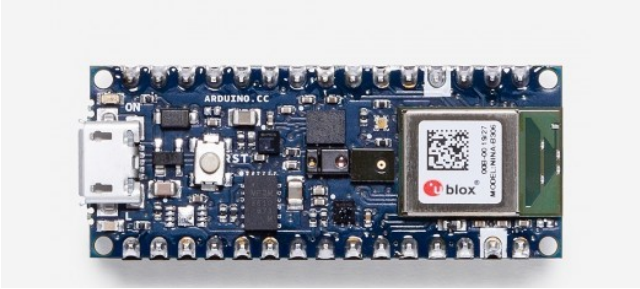
Privacy Benefits

- Data need not stored on servers i.e. less risks & opportunities for privacy leaks

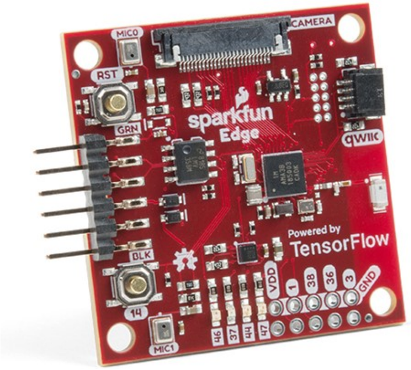
SAMPLE HARDWARE FOR TINYML

TinyML Research, Development and Experimentation

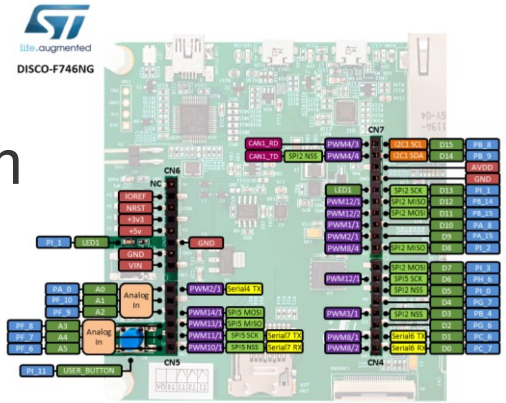
- ARDUINO NANO 33 BLE SENSE WITH HEADERS, Cost ~ 30€



- SparkFun Edge Development Board - Apollo3 Blue DEV-15170, Cost ~ 15€



- STM32F746G-DISCO discovery board (32F746GDISCOVERY) - complete platform for STMicroelectronics ARM® Cortex®-M7 core-based STM32F746NGH6 microcontroller, Cost ~ 70€



THE SOFTWARE INFRASTRUCTURE

How to Build ML Models for TinyML Environments

- Standard Tools Work for Data Science & Model Development:
  - Python, Jupyter Notebooks, Arduino IDE,…
- Machine Learning Framework for TinyML:
  - **<u>TensorFlow</u>**: Suite of tools for building and running ML models
  - **<u>Keras</u>**: TensorFlow's high-level API focused on building and training Deep Learning applications
  - **<u>TensorFlow Lite</u>**: Specifically designed for inference on devices with limited computing capacity (e.g., phones, tablets, embedded devices).
  - **<u>TensorFlow Lite Micro</u>**: Deploy models on microcontrollers and other devices with only few kilobytes of memory e.g.,
    - Core runtime just fits in 16 KB on an Arm Cortex M3
    - Doesn't require operating system support, any standard C or C++ libraries, or dynamic memory allocation

RESEARCH DIRECTIONS - 1

Going Beyond Simple TinyML

Optimizing ML Models and Applications across heterogeneous IoT devices

- Typical problem in IoT applications at the edge where many devices are involved
- Develop models for platforms with resource constraints
- Customize models to the needs of different platforms towards balancing accuracy and efficiency in the best possible way
- Train models for full kernel and gradually shrink in-line with the requirements of the hardware platforms involved

Integrate TinyML with AutoML

- AutoML: Automatically identify & deploy the best possible Neural Network Architecture taking into account device profiles and resources

RESEARCH DIRECTIONS - 2

Going Beyond Simple TinyML

Develop special ML pipelines and optimize them for specific use cases

- Optimization Criteria: Performance, Model size and Computation, as well as their trade-offs

Edge/Cloud Management platforms for TinyML devices and applications

- Management and deployment of large scale networks of IoT devices
- Workflows to train, build, deploy, and adapt TinyML models at scale

Federated Learning in Mobile Edge Networks with TinyML models

- Sharing model updates for improved performance
- Address heterogenous environments with TinyML nodes can be challenging due to diverse Communication costs, Resource Allocation requirements, as well as Privacy & Security Needs

EU-IOT ON TINYML

EU-IoT Plans for TinyML in collaboration with ICT-56 Projects

Including TinyML in the “Far Edge” Activities Research Roadmap

Develop Training Resources (Presentations, Webinars) on TinyML

Publish and Disseminate TinyML Success Stories

Aggregate and Publish the main Open Source TinyML Projects and Tools

Streamline with other Activities (e.g., Reference Architectures and Standards Development)

THANK YOU FOR YOUR ATTENTION

eu-iot.eu

EU-IoT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 956671